Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
http://br-ie.org/pub/index.php/rbie

Submission: 31/Mar/2022; 1st round notif.: 27/Jul/2022; New version: 07/Sep/2022; 2nd round notif.: 24/Sep/2022;
Camera ready: 26/Oct/2022; Edition review: 2/Nov/2022; Available online: 10/Nov/2022; Published: 10/Nov/2022;

Atribuição de autoria em trabalhos escolares por meio da estilometria e processamento de linguagem natural

Title: Authorship attribution in school works through stylometry and natural language processing

Daniel Cirne Vilas-Boas dos Santos
Centro de Informática - UFPE
ORCID:0000-0003-0665-151X
dcvs@cin.ufpe.br

Cleber Zanchettin
Centro de Informática - UFPE
ORCID:0000-0001-6421-9747
cz@cin.ufpe.br

Resumo

O aumento no volume de documentos digitais associado ao seu uso no processo de verificação de aprendizagem demanda recursos computacionais para compreensão e análise de autoria. A literatura propõe distinguir os autores pelo estilo de escrita e palavras-chave. Entretanto, estes trabalhos não estão inseridos no contexto educacional e são majoritariamente em inglês. Tal cenário é desafiador, pois apresenta menos documentos por autor, homogeneidade entre o grupo de autores e menor quantidade de trabalhos anteriores e recursos para o idioma. Devido ao baixo volume de exemplos, usamos bases jornalísticas robustas como referência. Por meio dos experimentos verificamos que, em domínios específicos, representações baseadas em características de estilo são superiores à abordagens textuais, as quais sofrem influência do tópico em corpora mais abrangentes. Este trabalho revelou que o comitê de árvores de decisão extremamente aleatórias associado às características de estilo propostas foi superior aos demais modelos em todas as bases utilizadas, alcançando uma média de 71% na taxa de acerto e AUC 0,81.

Palavras-chave: *Estilometria, Atribuição de Autoria, Classificação de Documentos Escolares, Extração de Características Estilométricas, Comitês de Árvores de Decisão*

Abstract

The growth of digital documents, associated with their usage in several knowledge areas requires computational resources for its comprehension and analysis. The literature proposes distinguishing authors by their writing style and keywords. However, these studies mainly involve journalistic and literary contexts written in English. This research is unique because it explores authorship analysis within a dataset composed of school activities written by undergraduate students in Portuguese. Such a scenario is challenging because it contains fewer documents per author, homogeneous authors, and fewer research and tools in Portuguese. Due to the insufficient number of samples, we used robust journalistic datasets as reference. The experiments verified that stylometric representations are superior to textual representations in restricted domains, which suffer from the topic's broader corpora. Furthermore, we found out that the ensemble of extremelly randomized decision trees associated with the proposed stylometric features overcome every other model tested, in allthe datasets, reaching an average accuracy of 0.71 and 0.81 AUC.

Keywords: *Stylometry, Authorship Attribution, Scholar Document Classification, Stylometric Feature Extraction, Decision Trees Ensembles*

Cite as: Dos Santos, D. C. V.-B., & Zanchettin, C. (2022). Atribuição de autoria em trabalhos escolares por meio da estilometria e processamento de linguagem natural. Revista Brasileira de Informática na Educação, 30, 648-671. DOI: 10.5753/rbie.2022.2588.

1 Introdução

A utilização de recursos tecnológicos na educação, apesar de trazer diversos benefícios, também possui mazelas inerentes a sua inserção. O número de práticas prejudiciais, como a reprodução de informações na web, comércio de trabalhos escolares e divisão paralela de atividades entre alunos têm se tornado mais frequentes (Singh & Remenyi, 2016; Curtis & Tremayne, 2019). Os recursos *online* são excelentes fontes para pesquisa, mas a compreensão e interpretação das informações de maneira individual é fundamental durante o processo de aprendizado. Segundo Werneck (2006), algumas práticas fundamentais para construção do conhecimento são a organização do pensamento, percepções de espaço, tempo, efeito e causa e visão de conjunto das partes. A facilidade de comprar trabalhos escolares ou reproduzir conteúdos da *web*, além de sedutora, é difícil de ser identificada durante a correção das atividades pelos avaliadores. Tais práticas atravessam e perturbam as etapas de construção do conhecimento citadas, e prejudicam todos os envolvidos no processo de ensino e aprendizagem. Dessa forma, a criação de recursos computacionais capazes de mitigar tais mazelas e apoiar o processo de verificação da aprendizagem se faz necessária.

Uma das formas de minimizar esse problema, considerando que a modalidade mais comum para realização de atividades pedagógicas é por meio da comunicação escrita, é a análise de autoria, a qual consiste em uma atividade dentro da classificação de documentos focada nos autores. Esta atividade tem se apresentado como uma estratégia eficiente para resolução de problemas semelhantes em outros contextos ((Tempestt et al., 2017). Os principais desafios dentro do problema de pesquisa são, *i*) o pequeno número de documentos por autor, pois é provável que só exista um volume significativo de documentos por estudante após algum tempo de curso; *ii*) domínio restrito ao conteúdo do curso ou disciplina, que colabora com a presença de muitos documentos semelhantes; e *iii*) uso indiscriminado de ferramentas de busca, que leva à construção de textos compostos por excertos de outros autores, dificultando a análise do discurso e na identificação do estilo de escrita dos estudantes.

Este trabalho é uma versão estendida do artigo premiado no XXXII SBIE: Simpósio Brasileiro de Informática na Educação (Santos & Zanchettin, 2021), que explora a atribuição de autoria em atividades escolares na língua portuguesa, visando sustentar o crescimento da tecnologia na educação e desencorajar práticas nocivas durante o processo de construção e verificação da aprendizagem de maneira não punitiva (Botelho & da Silva Martins, 2020). Após a condução de um estudo de caso que avaliou diversas abordagens para resolução da atividade, utilizou-se o Processamento de Linguagem Natural (PLN) para pré-processamento e extração de características de estilo na língua portuguesa associado a comitês de árvores de decisão extremamente aleatórias (Geurts, Ernst & Wehenkel, 2006).

As etapas do estudo estão apresentadas nas próximas quatro seções. A seguir, discutem-se as bases teóricas para a investigação da análise de autoria, estilometria e aplicação de PLN e Árvores de Decisão. Na seção 3, há o detalhamento da metodologia, que inclui o desenvolvimento da extração de características estilométricas. A quarta seção detalha os experimentos executados e seus respectivos resultados. Por fim, estão elencadas as conclusões e perspectivas de desenvolvimento de trabalhos futuros.

2 Fundamentação

A análise de autoria é uma área de estudo voltada a solucionar problemas relacionados à autoria de documentos. Seu crescimento está vinculado à ampliação do uso da computação forense para solução de problemas cotidianos. Na literatura, encontramos trabalhos dedicados à identificação de *fake news* (Peng, Choo & Ashman, 2016), combate ao plágio (Curtis & Tremayne, 2019), identificação de pseudônimos, práticas abusivas online (Gillam & Vartapetiance, 2012) e solução de casos com autoria contestada ou anônima (Custódio & Paraboni, 2021). Além disso, o desenvolvimento da Aprendizagem de Máquina (AM) e do Processamento de Linguagem Natural (PLN) contribuíram para incrementar as pesquisas da área – AM e PLN são ferramentas essenciais para a compreensão dos documentos e construção dos modelos de classificação.

Dentro desse campo, se destacam as tarefas de Atribuição de Autoria, Verificação de Autoria e Caracterização de Autoria, as quais, respectivamente, identificam, confirmam e caracterizam os autores dos documentos a partir do seu conteúdo (Juola, 2008) (Stamatatos, 2009).

A estilometria defende o uso de características de estilo para quantificar e definir o estilo de escrita dos autores. Segundo Stamatatos (2009) e Varela, Albonico, Justino, Bortolozzi et al. (2018), cada autor possui um estilo único de escrita, que é composto por múltiplos fatores, como vícios de linguagem, uso e composição de pontuação, palavras, frases e parágrafos, legibilidade, concordância e riqueza de vocabulário. Porém, a produção de tais características exige esforço humano para extração, construção e avaliação. A seleção de características é apontada como um dos maiores desafios da estilometria, pois não há consenso sobre quais são mais relevantes, e o conjunto delas pode variar bastante de acordo com o problema (Neal et al., 2017).

Para extração dessas características, faz-se necessária a compreensão da linguagem a nível de *tokens*, frases e parágrafos. Isso pode ser alcançado por atividades do PLN, baseadas em corpus, sistemas de referência léxica, expressões regulares ou análise e geração de regras gramaticais (Chowdhury, 2003). Exemplos destes são os anotadores sintáticos, classificadores de entidades nomeadas e o *Chunking*. A anotação sintática é feita por meio de um classificador (*POS-Tagger*), capaz de etiquetar gramaticalmente elementos textuais no documento (Kumawat & Jain, 2015). Os classificadores de entidades nomeadas funcionam de maneira análoga, e são capazes de identificar as entidades (locais, pessoas, instituições). Ambos os classificadores são treinados por meio de aprendizagem supervisionada e bases rotuladas (corpora). O *Chunking* é uma técnica aplicada para facilitar a compreensão das frases a nível computacional por meio da identificação dos principais constituintes das frases. Esta técnica é geralmente associada à anotação sintática e semântica para realização de casamento de padrões e extração de sequências considerando regras gramaticais e expressões regulares (Ramshaw & Marcus, 1999).

Além de estratégias baseadas no PLN, existem métodos para representação do documentos diretamente baseados nas palavras, como os vetores de contagem de palavras por TF-IDF (*Term Frequency–Inverse Document Frequency*) e *Bag of Words* (Goldberg, 2017). Através da transformação dos documentos textuais em matrizes numéricas, é possível alimentar os modelos de AM.

As árvores de decisão (AD), a exemplo, são um conjunto de algoritmos de AM representados em uma estrutura de árvore que é composta por nós, na qual cada um pode representar uma classe (nó de resposta), ou uma condição de teste para partição dos elementos restantes. As carac-

terísticas dos exemplos de entrada são avaliadas em cada nó de condição, até que seja alcançado algum nó de resposta. Geralmente são adotados critérios de minimização de erro em nós com mais de uma opção de classe. Na Figura 1 temos uma ilustração de AD que apresenta a partição dos exemplos a partir dos critérios de seleção. Critérios de seleção com maior ganho de informação (menor entropia) se localizam mais próximos a raiz pois têm maior capacidade de separação de exemplos em classes.

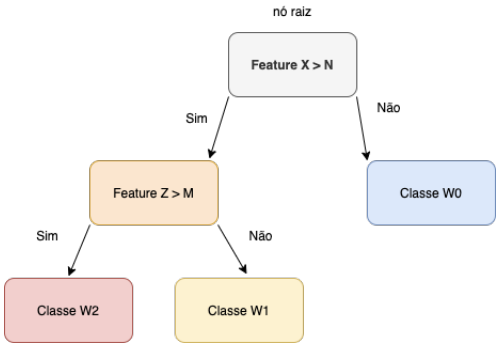

Figura 1: Ilustração de uma árvore de decisão com nós de condição e resposta. No exemplo a árvore classifica os exemplos nas classes W0, W1 e W2 a partir das características X e Z..

Nos últimos anos, foram desenvolvidas técnicas de otimização sobre ADs em busca de classificadores mais eficientes ao associá-las com *bagging* (Breiman, 1996), *boosting* (Schapire, Singer & Singhal, 1998), camadas de aleatoriedade e comitês de classificadores. Os comitês (*ensembles*) são representados pela associação de vários classificadores, cada um destes realiza suas predições e a combinação das predições individuais representa a decisão do comitê. Esta decisão geralmente é feita por meio de algum critério, como o voto majoritário ou ponderado (Kocev, Vens, Struyf & Džeroski, 2013).

Dentre os comitês de ADs mais tradicionais, destacamos a Floresta Aleatória (*Random Forest* - RF), Árvores Extremamente aleatórias (*Extremelly randomized Trees* - ET) e Árvores de Decisão impulsionadas com aumento de gradiente extremo (*Extremelly Gradient Boosted Trees* - XGBT). O RF é um comitê de ADs que gera uma diversidade de árvores associadas por meio de *bagging* e adiciona camadas extra de aleatoriedade por meio de subconjuntos da base (*bootstrap*), que ajuda na generalização (Breiman, 2001). O ET também é um *ensemble* de ADs que constrói árvores com pontos de corte aleatórios a partir de toda a base de dados, gerando um conjunto mais robusto de árvores profundas, prevenindo o sobreajuste (Geurts et al., 2006). Já o XGBT é uma versão mais aprimorada das Árvores com aumento de gradiente, que se assemelha ao RF e ET, porém também inclui uma impulsão nos resultados através do aprendizado por reforço, realizado a partir de predições anteriores incorretas e controlada por uma função de perda (Chen & Guestrin, 2016) (Baker, Isotani & Carvalho, 2011).

A associação de modelos baseados em AD a características de estilo tem resultados comprovados na literatura para atividades de atribuição e verificação de autoria (Khonji, Iraqi & Jones, 2015; Pacheco, Fernandes & Porco, 2015; Maitra, Ghosh & Das, 2016). Também vale ressaltar que modelos do tipo máquina de vetores de suporte (*Support Vector Machines* - SVM) - redes neurais recorrentes e aprendizagem não supervisionada - vêm sendo empregados junto a características textuais e estilométricas para atividades de atribuição e caracterização de autoria (Yang et al., 2018; Bevendorff et al., 2020).

3 Metodologia

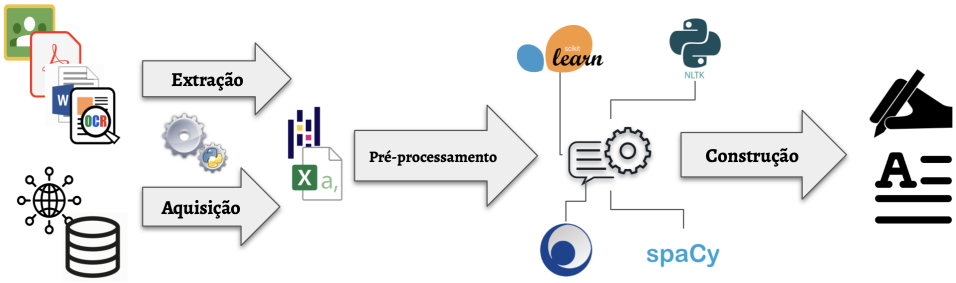

Figura 2: Etapas para construção das bases de dados. Após a extração e aquisição dos dados brutos, foram realizados o pré-processamento, a construção e extração de características e a representação destes utilizando as abordagens estilométrica e textual.

A Figura 2 detalha a metodologia adotada para a pesquisa. O processo se iniciou com a extração e aquisição dos dados brutos, com subsequente realização do pré-processamento (mineração) destes dados. Mais adiante foram implementados mecanismos para cálculo e extração de características, tomando como base a literatura e a experimentação. Por fim, esses valores foram transformados em bases de dados representadas de maneira estilométrica e textual. Nas subseções a seguir detalhamos cada uma das etapas ordenadamente.

3.1 Extração de dados

Apesar da existência de bases de dados relevantes voltadas para atividades de atribuição de autoria (Varela et al., 2018; Bevendorff et al., 2020), nenhuma atendeu por completo os requisitos desta pesquisa. Para investigar o problema proposto, faz-se necessário obter documentos escritos por estudantes na língua portuguesa. Bases como a *Student Performance*(Cortez & Silva, 2008) percorrem o mesmo grupo de estudo, mas condensam apenas dados geográficos e comportamentais para correlacionar com o desempenho final dos estudantes. Outros conjuntos de dados, como CAPES (Soares, Yamashita & Anzanello, 2018) e CorpusTCC(Pardo & Nunes, 2003) agrupam documentos escritos em português no contexto acadêmico, mas são majoritariamente compostos por autores com apenas um documento, o que dificulta estudos sobre análise de autoria

Desta forma, por meio de uma parceria com professores de ensino superior, construímos uma base composta por atividades pedagógicas escritas por estudantes (Figura 3). Ao remover trabalhos realizados em grupo ou com pelo menos 3 exemplos por autor, a base foi reduzida a 84 documentos distribuídos entre 16 autores. Dado o baixo volume de exemplos e grande disparidade no número de trabalhos e *tokens* por autor na base de **estudantes**, utilizamos outras duas bases para efeitos comparativos (Tabela 1). **notícias**: base composta por textos jornalísticos extraídos de um renomado portal de notícias brasileiro[1] no ano de 2021. Esta base foi construída durante a pesquisa de forma manual. Realizamos a extração do título e corpo das matérias, que foram associados à uma classe (colunista). A intenção foi construir uma base com estrutura semelhante à base de estudantes (número de documentos e autores), porém com um balanceamento intencional no número de documentos por autor, e **Varela**: coleção robusta de textos jornalísticos,

[1]Estadão - https://www.estadao.com.br

escritos por diversos autores e distribuídos por assunto em 10 categorias distintas (Varela et al., 2018). Na tabela abaixo detalhamos as bases a nível de exemplos, autores, *tokens* e a razão de desbalanceamento[2] (Tarekegn, Giacobini & Michalak, 2021).

Tabela 1: Descrição das bases de dados.

Base	Documentos	Tokens	Autores	Proporção máxima de desequilíbrio
Estudantes	84	60.721	16	3,33% (Desbalanceada)
Notícias	200	125.998	10	1% (Balanceada)
Varela	3.000	1.426.044	100	1% (Balanceada)

A segunda Conferência organizada pela UNESCO e PNUMA foi a Conferência de Tbilisi, a primeira e mais importante Conferência sobre Educação Ambiental a nível intergovernamental. Endossada por 150 países dentre os quais o Brasil não participou em caráter oficial. Além disso, foi crucial para o desenvolvimento da primeira fase do Programa Internacional de Educação Ambiental (PIEA), que foi inicialmente sugerido na Conferência de Estocolmo e realmente iniciado somente na Conferência de Belgrado. Ou seja, a relevância do evento promovido durante treze dias no período de 14 a 26 de outubro de 1977 teve seu ponto chave o estabelecimento do PIEA a qual foi sugerido em Estocolmo (1972) e Belgrado (1975).

Nesse sentido durante o evento foram organizadas quarenta e uma recomendações sobre educação ambiental a nível mundial, considerados um grande marco na educação ambiental. Nesse sentido DIAS (2000) ressalta a relevância presente nas ações do documento deixa claro que a educação ambiental deve considerar não somente a fauna e a flora, mas incluir também os aspectos sociais, econômicos, científicos, tecnológicos, culturais, ecológicos e éticos. Em seus objetivos: a) consciência: ajudar os grupos sociais e os indivíduos a adquirirem consciência do meio ambiente global e ajudar-lhes a sensibilizarem-se por essas questões; b) conhecimento: ajudar os grupos e os indivíduos a adquirirem diversidade de experiências e compreensão fundamental do meio ambiente e dos problemas anexos; c) comportamento: ajudar os grupos sociais e os indivíduos a comprometerem-se com uma série de valores, e a sentirem interesse e preocupação pelo meio ambiente, motivando-os de tal modo que possam participar ativamente da melhoria e da proteção do meio ambiente; d) habilidades: ajudar os grupos sociais e os indivíduos a adquirirem as habilidades necessárias para determinar e resolver os problemas ambientais; e) participação: proporcionar aos grupos sociais e aos indivíduos a possibilidade de participarem ativamente nas tarefas que têm por objetivo resolver os problemas ambientais.

Figura 3: Documento de exemplo da base de dados. Atividade escrita por um estudante da disciplina de Educação Ambiental.

3.2 Pré processamento

Durante o pré-processamento das bases, garantiu-se o anonimato e eliminou-se vieses por meio da remoção de termos e palavras-chave que pudessem estar vinculados à autoria. Para tanto, foi feita a implementação de um *script* de limpeza de nomes próprios (alunos, escritores e professores), datas de escrita dos documentos, nome da disciplina, instituição de ensino e título das atividades propostas.

3.3 Extração de características

Para extração de características, fizemos uso massivo de recursos de AM e PLN. As características extraídas foram separadas em grupos lógicos conhecidos na literatura e descritos abaixo (Tabela 4.2.3).

[2] *Maximum Imbalance Ratio* - Razão entre as classes majoritária e minoritária

1. **Lexicais**: representam a estrutura da escrita dos autores, diretamente relacionada com a distribuição das letras, palavras, sentenças e parágrafos ao longo do documento. Neste grupo, encontramos características relacionadas à frequência e tamanho de parágrafos, frases e sílabas por palavra. Combinamos o algoritmo de separação de sílabas proposto por (Silva, 2011) com a implementação da biblioteca *Pyphen*[3] para obter uma separação mais precisa.

2. **Caracteres e palavras-chave**: representam a frequência de aparição de palavras ou pontuação ao longo do texto. No grupo de termos pré-definidos, foram mensuradas as pontuações menos comuns (e.g. ponto e vírgula, dois pontos, travessão), conectivos lógicos (e.g. e, ou, desde que) e palavras capitalizadas. Também se calculou a frequência de aparições dos *n-grams* ($n = [2, ..., 5]$) mais frequentes do *corpora* em cada documento (*top-grams*);

3. **Sintáticas**: indicam o papel morfo-sintático dos *tokens* e frases no documento. Para anotação sintática, foi treinado um *POS-Tagger* na língua portuguesa sobre a base MACMORPHO (Aluísio et al., 2003). Desta forma, obtivemos a frequência de classes gramaticais para criação de características. Na seção anotação sintática da Figura 4 temos um texto anotado sintaticamente. Cada uma das palavras é associada à sua classe gramatical por meio de uma cor e de texto em negrito.

   Além das próprias características mensuradas diretamente pelo anotador sintático, este processo permitiu a identificação de sintagmas verbais e nominais por meio do *chunking* associado às expressões regulares abaixo (Duarte, 2021).

$$SN = (ARTIGO^{?} ADJETIVO^{*} (SUBSTANTIVO|PRONOME)^{+})$$

$$SV = (VERBO)^{?} ADVÉRBIO^{*} VERBO^{+})$$

   Após a separação do documento em frases, foi realizado um casamento de padrões para mensurar a incidência dos sintagmas. Na seção estrutura frasal da Figura 4 temos um exemplos de sintagmas extraídos a partir da base de estudantes.

   Outras duas características que compõem este grupo oriundas da anotação sintática são a frequência de palavras de conteúdo e as palavras funcionais, pelas quais as palavras são categorizadas de acordo com sua classe gramatical (Scarton & Aluísio, 2010).

   Por fim, características sintáticas mais detalhadas, como flexões de gênero, plural, pessoas do discurso e tempo verbal também foram mensuradas e se encontram neste grupo. Para extrair tais características, foram realizadas adaptações no anotador sintático pré-treinado que se encontra disponível na biblioteca SpaCy (Honnibal & Montani, 2017).

4. **Semânticas**: quantificam os diversos papéis semânticos das palavras de acordo com o contexto em que estão inseridas. Este grupo está intimamente relacionado ao grupo sintático, pois a entrada dos classificadores de entidade nomeada são termos anotados sintaticamente. Para essa identificação, utilizou-se um classificador pré-treinado para reconhecimento de entidades nomeadas. Tal classificador foi disponibilizado no trabalho de (Pires, 2017) e se baseia na base de dados HAREM (Freitas et al., 2010). Ele é capaz de rotular termos de acordo com categorias e subcategorias de entidades do próprio HAREM. Foram mensuradas todas 10 categorias de entidades, além das proporções entre todos os termos rotulados

[3]Pyphen (https://pyphen.org) é um módulo para hifenizar palavras usando dicionários de hifenização

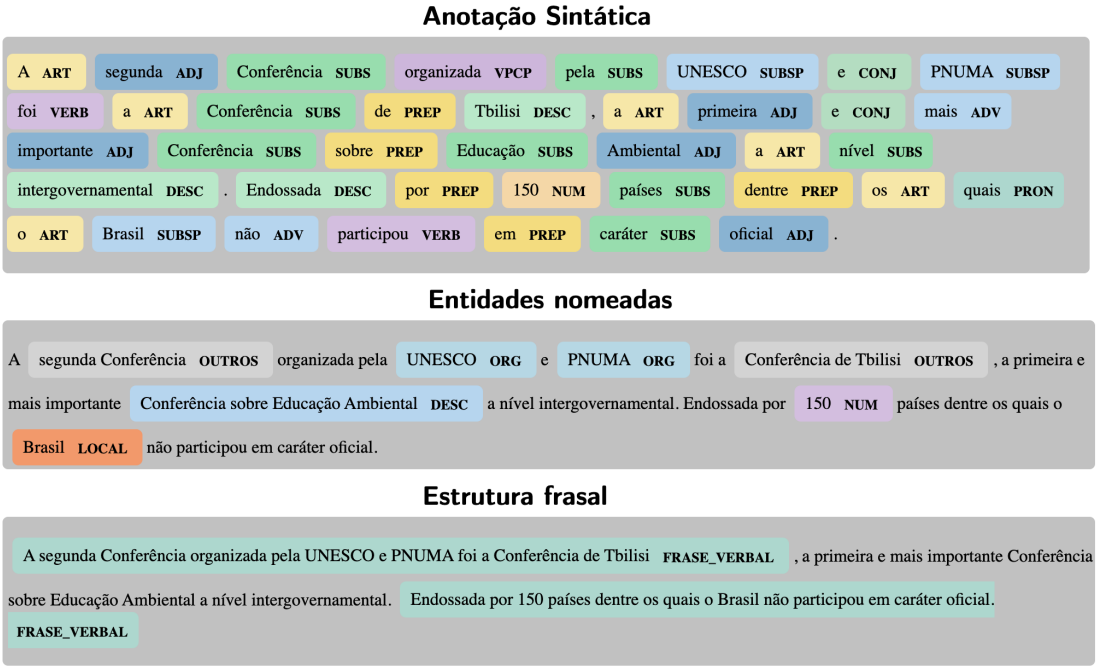

Figura 4: Ilustração da aplicação de PLN para categorização de palavras e frases. Esse processo serviu de base para extração de características estilométricas dos grupos sintático e semântico. A imagem acima foi gerada utilizando o visualizador da biblioteca SpaCy .

e aqueles não rotulados dentro de alguma das categorias de acordo com o classificador pré-treinado. Na seção de entidades nomeadas da Figura 4 são exibidos exemplos de palavras classificadas nos grupos de entidades organização, local, valores numéricos e outros.

5. **Riqueza de vocabulário**: esse grupo contém índices que indicam a legibilidade e diversidade lexical dos textos. Ele contém características computadas através da frequência de termos únicos (*hapax legomena*) considerando cada documento (local) ou todo o *corpora* (global), repetição de palavras, legibilidade - calculada por meio do índice Flesch-Kincaid, adaptado para o português (Martins, Ghiraldelo, Nunes, de Oliveira Junior et al., 1996)e ; diversos outros coeficientes de riqueza de vocabulário (Tweedie & Baayen, 1998), como as medidas K (Yule, 1944), H (Honoré, 1979) e U (Dugast, 1979).

6. **Aplicação**: características relativas ao domínio da aplicação, geralmente baseadas em listas de palavras e dicionários. Neste grupo, foram consideradas as listas de *stopwords* e *collocations* ($n = [2, ..., 4]$) do NLTK e um dicionário ortográfico.

3.4 Representação das bases

Conforme já mencionado, foram utilizadas as representações textuais e numéricas para retratar os documentos. Na representação textual, aplicou-se TF-IDF e *word-embeddings* para transformação dos dados. Os *word-embeddings* aplicados foram o Glove, FastText e Word2Vec (Bojanowski, Grave, Joulin & Mikolov, 2017; Jang, Kim & Kim, 2019), com dimensões de tamanho 50 e 100. Na representação numérica, empregam-se as características estilométricas detalhadas na Subseção 3.3. Esta representação da base é composta pelas 74 características de estilo extraídas

Tabela 2: Descrição dos grupos de características.

Grupo (Quantidade)	Detalhamento
Lexical (8)	Estrutura do documento em relação aos parágrafos, frases, palavras únicas e sílabas
Caracteres ou palavras-chave (13)	Frequência de palavras, pontuações e *top grams*
Sintático (25)	Classes gramaticais, flexões de gênero, plural e tempo verbal
Semântico (11)	Entidades nomeadas
Riqueza de vocabulário e legibilidade (12)	Índices de riqueza de vocabulário e legibilidade
Específico da aplicação (5)	Baseados em listas de palavras, como dicionários ortográficos, *stopwords* e *collocations*

e computadas por meio de um algoritmo desenvolvido nesta pesquisa. Tal algoritmo é capaz de extrair os dados a partir de fontes em diversos formatos, realizar o pré-processamento, calcular e exportar os dados nos formatos textual e estilométrico das bases estudadas a partir dos documentos originais[4].

4 Experimentos e discussões

Os experimentos foram divididos nas seguintes fases: 1) análise exploratória e agrupamento, 2) avaliação de classificadores, 3) otimização dos modelos, e 4) interpretação dos resultados.

4.1 Análise exploratória e agrupamento

Inicialmente, inspecionou-se a base de estudantes a nível de palavras e exemplos por autor. Foram realizadas observações considerando a autoria (*tokens* por autor) e considerando os autores isoladamente (*tokens* por autor por documento). Considerando apenas a autoria, temos um total de 46.477 palavras distribuídas entre 84 trabalhos, contabilizando uma média ($\overline{x}$) de 553,29 *tokens* por documento. A alta variabilidade de *tokens* por autor é confirmada pelo desvio padrão (σ) de 782,08. Já na comparação por autores, documentos e *tokens*, observamos uma diferença muito grande para cada um dos autores. Por exemplo, a maior média foi do autor "0", com 8655 *tokens* distribuídos entre 4 documentos de sua autoria. Com isso, a média de *tokens* por documento deste autor foi de $8655/4 = 2163,75$, que é a maior média de entre todos os estudantes. Do outro lado, temos a menor média que é do autor "4" com apenas 756 *tokens*, uma média de $\overline{x} = 189$ *tokens* por documento com desvio padrão $\sigma = 503,52$. De maneira geral, o desvio padrão do número de *tokens* por autor é ainda elevado ($\sigma = 2143,41$), valor que confirma a discrepância de *tokens* por autor.

Esta análise ratificou o desbalanceamento da base, tanto a nível de exemplos como a nível de palavras por autor. Em conversas com professores parceiros para entender as razões dessa disparidade, foram citadas justificativas como: *i*) alguns alunos desistiram de determinadas disciplinas

[4]O algoritmo produzido durante esta pesquisa foi disponibilizado sob licença de código aberto. Disponível em: https://github.com/daanielvb/text-extractor

ou do curso; *ii*) alguns alunos preferiam ou só podiam entregar os trabalhos escritos à mão; *iii*) algumas das atividades eram complementares ou opcionais; *iv*) algumas disciplinas tiveram mais atividades escritas que outras e; *v*) alguns alunos se dedicaram mais às atividades que outros.

Após estas primeiras constatações com análise das métricas, partiu-se para a criação de visualizações interpretáveis de todas as bases, através da combinação das técnicas de redução de dimensionalidade PCA e TSNE (Van der Maaten & Hinton, 2008).

Na Figura 5 temos os exemplos das bases de dados representados em duas dimensões. Cada sub-gráfico é uma combinação entre a base de dados e a técnica para representação dos dados. As cores representam a classe dos exemplos (autoria). Na imagem, é possível observar que as fronteiras de separabilidade dos exemplos por autoria na base de estudantes era muito pequena, tanto nas representações textuais como de estilo. Por outro lado, nas bases jornalísticas a separabilidade e segregação são mais evidentes. Os principais conglomerados de exemplos de um mesmo autor numa única região do gráfico foram marcadas para fins ilustrativos.

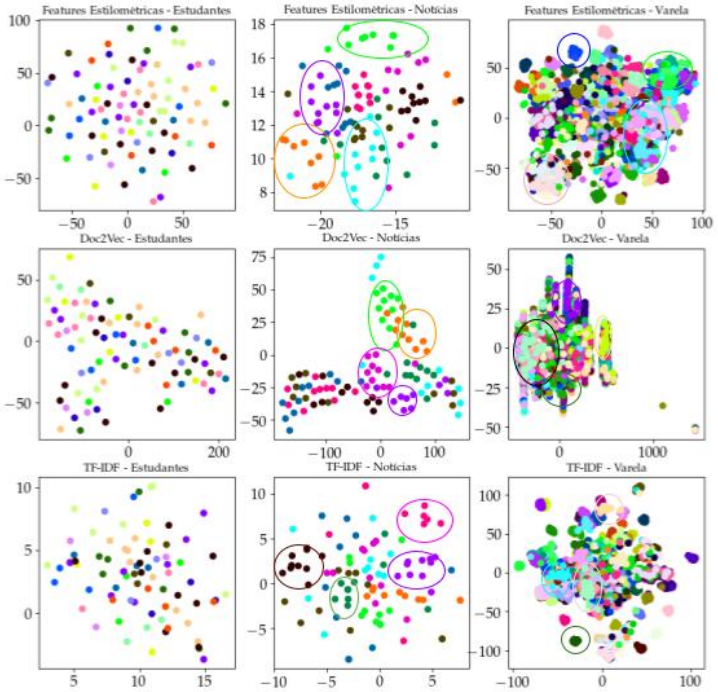

Figura 5: Apresentação das bases de dados em 2 dimensões após redução de dimensionalidade com PCA + TSNE. É possível observar na primeira coluna que a base de estudantes não apresenta regiões com segregação de exemplos por autor, diferentemente das outras bases dispostas na segunda e terceira coluna.

Como parte da análise exploratória, foi conduzida uma análise de *cluster* com propósito de avaliar possíveis agrupamentos a partir da distribuição dos dados. Foram aplicados dois algoritmos de agrupamento, um com partição suave (*Fuzzy* c-médias) e outro com partição dura (K-médias). Os valores de K foram definidos de duas maneiras, *i*) a partir do conhecimento prévio sobre a base (K = número de autores para base) e *ii*) na distribuição dos exemplos, por meio do cálculo da silhueta (Thinsungnoena et al., 2015) ou índice de partição difusa (FPC) (Bezdek, 2013).

Para a base de estudantes, ao inspecionar os agrupamentos construídos a nível de autoria e palavras, observou-se majoritariamente que as atividades dos estudantes apresentavam maior coesão quando tratavam da mesma atividade pedagógica ou atividades da mesma disciplina (escritas

por autores distintos). Ainda assim, foi comum observar palavras repetidas, como "ensino", "ciência"e "conhecimento", em agrupamentos distintos, o que não foi tão comum nos agrupamentos gerados para a base de Varela (Figura 6). Com relação à atribuição de autoria, o agrupamento por meio das técnicas mencionadas não demonstrou ser eficaz para esta base.

Figura 6: Nuvens de palavras oriundas dos agrupamentos gerados por meio do K-médias com K derivado da silhueta. Na imagem temos uma nuvem para cada combinação de base e representação (estilométrica e textual) extraída de algum dos agrupamentos (c). Aqui observamos a repetição de palavras-chave relacionadas aos temas principais de cada agrupamento.

Para as bases jornalísticas, observamos agrupamentos capazes de segregar a maioria das obras de alguns autores. Nos experimentos a partir da representação textual, destacamos que vários agrupamentos demonstraram ser ocasionados em razão do tema do documento (Figura 7).

O fato de os autores na base Varela terem escrito sobre apenas um único tema levantou preocupações, pois os classificadores podem apresentar viés. Durante o agrupamento, especialmente na representação textual, observou-se que grande parte dos exemplos se aglutinou de acordo com o assunto. Considerando as palavras mais frequentes destes agrupamentos, percebeu-se, neste caso, a separação muito mais fácil, pois haviam termos únicos por agrupamento e poucos termos repetidos em agrupamentos distintos. Ou seja, uma alta coesão de exemplos por tema/assunto.

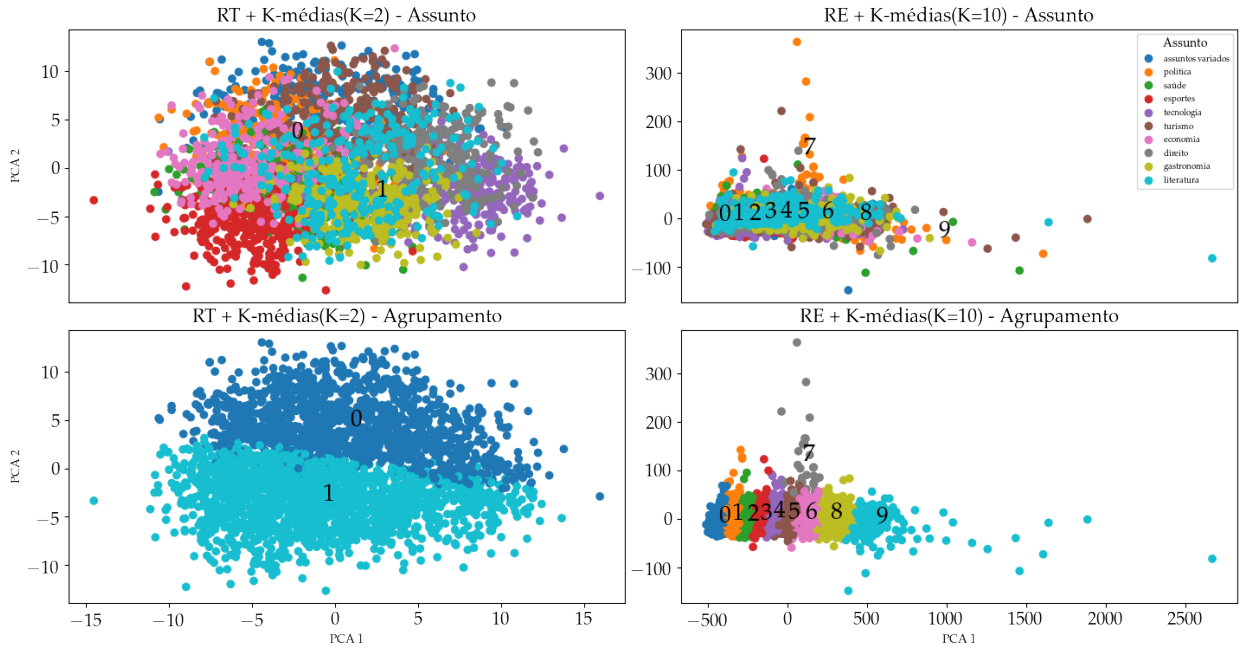

Figura 7: Distribuição dos dados na base Varela agrupados por assunto utilizando valores de K oriundos da silhueta e classes (autoria). É possível observar que o tema dos documentos exerce forte influência sobre os agrupamentos gerados e que a técnica não é capaz de separar documentos por autoria..

4.2 Avaliação, seleção e otimização dos modelos

4.2.1 Seleção

Os exemplos foram separados em treinamento e teste, respeitando a proporção 70/30% de maneira estratificada, em seguida foram fornecidos a uma série de classificadores reconhecidos na literatura (SVM, Regressão Linear, Naive Bayes, *Random Forest*, *Extremelly randomized Trees*, Árvores com aumento de gradiente , Perceptron Multicamada (MLP), Redes Neurais customizadas, Redes Neurais Convolucionais e LSTM). Utilizamos os parâmetros *default* das bibliotecas *Sklearn*, *Tensorflow* e *Keras* como ponto de partida. Os classificadores foram avaliados pela acurácia e área abaixo da curva (AUC), levando-se em conta o desbalanceamento.

Na Figura 8, é possível visualizar os primeiros resultados dos experimentos com classificadores tradicionais. Os classificadores que não superaram os modelos de base de referência em ao menos uma das representações ou *dataset* foram removidos. Para as bases jornalísticas, usamos como base de referência um ponto de corte a partir dos valores mínimos dispostos na literatura para a base Varela (60% de acurácia e 70% de AUC). Para a base de estudantes, como os resultados iniciais ficaram abaixo do ponto de corte, foram removidos os classificadores com desempenho inferior (último quartil) em relação aos demais.

Apesar dos resultados positivos na literatura o uso de *Deep Learning* combinado a *embeddings* pré-treinados (Shrestha et al., 2017; Bojanowski et al., 2017; Jang et al., 2019), o modelo não alcançou resultados satisfatórios, especialmente na base de estudantes. Isso pode estar relacionado ao baixo volume de exemplos e à perda de palavras importantes, ausentes nos *embeddings*. Foram utilizados *embeddings* treinados para o português com o *corpora* STIC 2017 (Hartmann et al., 2017) com dimensionalidades de tamanho 50, 100 e 300. Não foi observada diferença signi-

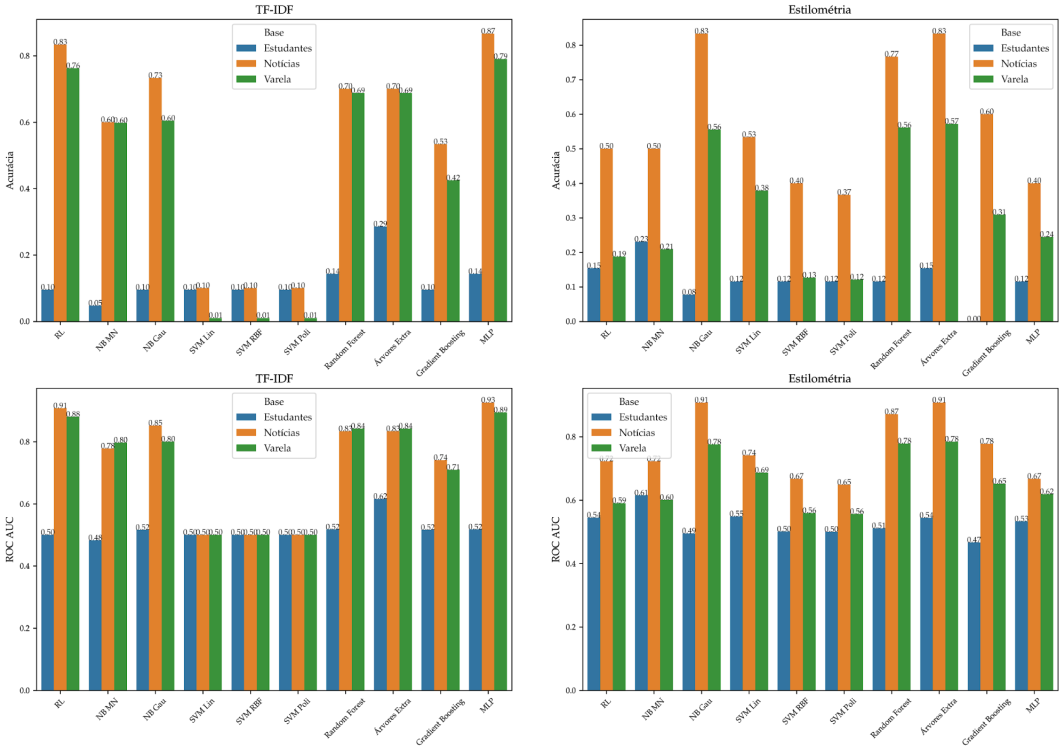

Figura 8: Resultados preliminares obtidos através da combinação de modelos tradicionais de aprendizagem de máquina e técnicas de representação (TF-IDF e estilométrica) sem otimização de parâmetros nem normalização de dados.

ficativa na acurácia dos classificadores ao variar a dimensionalidade. Dentre as estratégias para construção dos *embeddings*, o FastText utilizando *Skip-gram* mostrou resultados superiores aos demais (Mikolov, Chen, Corrado & Dean, 2013).

Outro resultado negativo que destacamos foram os modelos de máquina de vetor de suporte (SVM), que foram todos eliminados nesta etapa. Esse efeito pode estar relacionado com a ausência de técnicas para normalização dos dados e otimização dos hiperparâmetros, artifícios que favorecem esse tipo de classificador (Agarap, 2018).

4.2.2 Otimização

Diante de uma quantidade menor de modelos, experimentou-se aplicar técnicas de mudança de escala (escalonamento padrão, normalização mínimo-máximo e *Power Transformer* (Weisberg, 2001)) e otimização de hiperparâmetros. Para essa etapa, realizamos a avaliação e seleção dos melhores modelos por meio da validação cruzada estratificada com 3 *folds* por 10 iterações (dado o número mínimo de documentos por autor).

Para otimizar os parâmetros, fez-se uma combinação aleatória entre os possíveis valores para os parâmetros de cada modelo, limitada a até 50 opções, seguido por uma busca em *grid* limitada. Isso significa que para cada modelo em questão, foram treinados 50 outros modelos com parâmetros distintos. Ao final deste processo, selecionamos o melhor modelo dentre os 50 e comparamos com o modelo de base de referência (sem ajuste de parâmetros). Nos casos de melhoria, substituímos o modelo de base de referência pelo otimizado.

Os modelos baseados na representação de estilo apresentaram maiores ganhos do que os textuais, assim como classificadores pautados em Árvores de Decisão e Redes Neurais se comparados aos classificadores probabilísticos. O escalonamento padrão alcançou maiores ganhos na representação textual e a normalização mínimo-máximo teve o melhor desempenho na representação estilométrica.

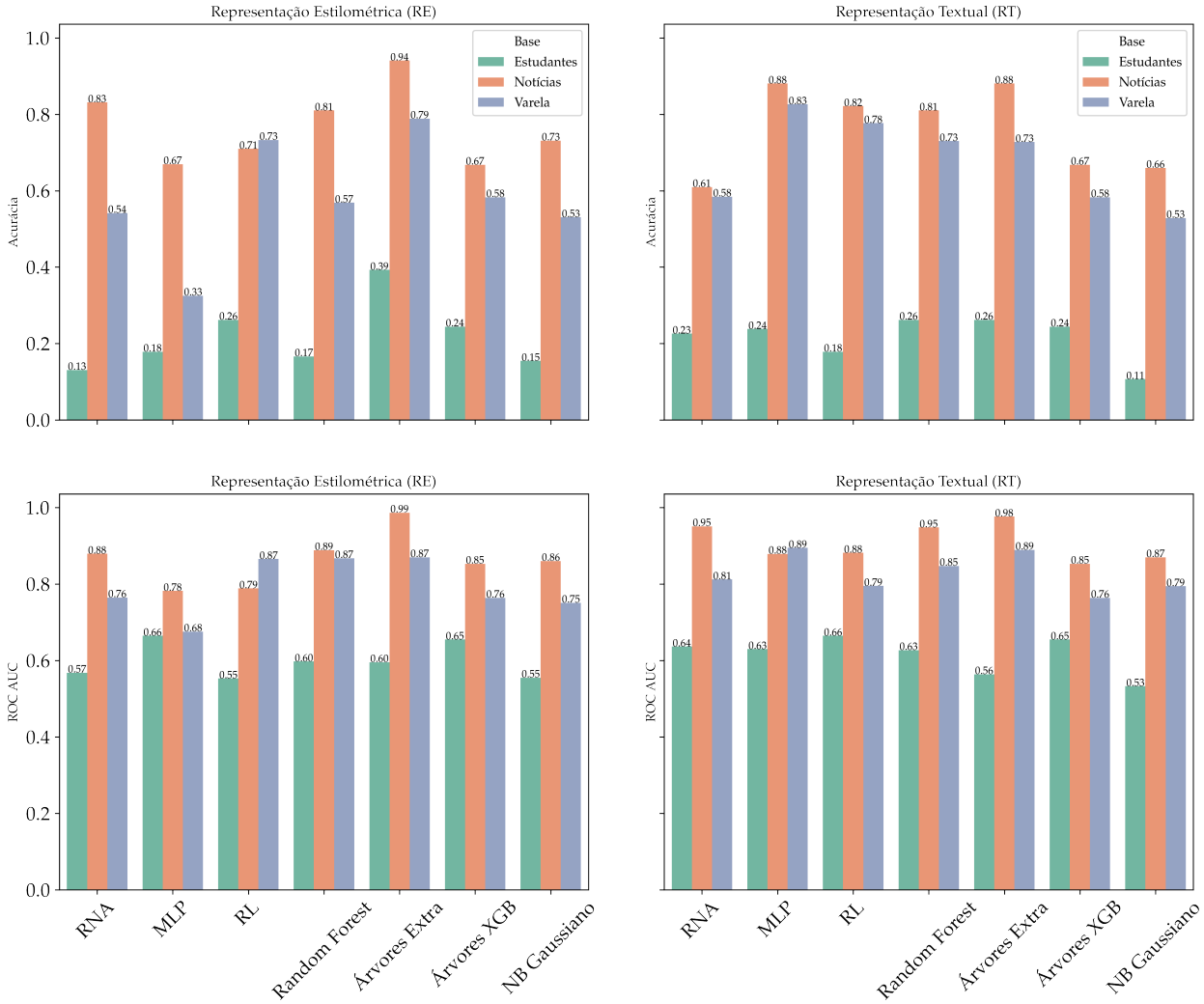

Figura 9: Resultados experimentais pós otimização agrupados por métrica, modelo e base. Observa-se que os modelos baseados em comitês de Árvores de Decisão e MLP apresentaram valores elevados para acurácia e AUC, especialmente para as bases jornalísticas.

4.2.3 Avaliação

As dificuldades previstas para a base de estudantes se confirmaram por meio dos resultados finais (Figura 9). O melhor classificador para esta base foi o comitê de Árvores Extra (*Extra Trees* - ET) na representação de estilo com 39% de acurácia e 0.60 AUC (Ainda superando um classificador aleatório com 6.25%). Na base de notícias, os resultados chegaram a 94% de acurácia e 0.99 AUC com um modelo da mesma composição. Para Varela, a MLP textual alcançou o maior resultado, com 83% de acurácia de 0.89 AUC, seguida pelo ET, também na representação de estilo. No critério geral, considerando as três bases de dados, o *ranking* (Figura 10) confirma que o ET

na representação de estilo foi superior aos demais classificadores, alcançando 71% de acurácia média.

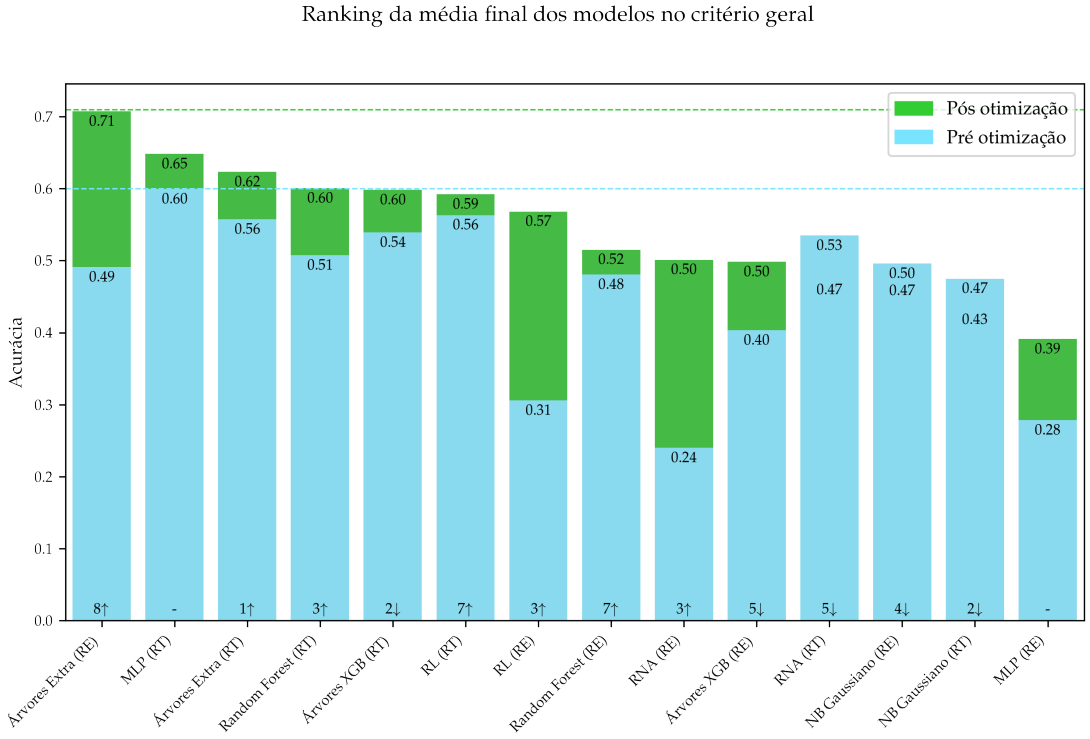

Figura 10: Acurácia média pré e pós otimização agrupados por classificador e representação. O ranking ilustra que os classificadores baseados em AD predominam entre os melhores.

Para ratificar os resultados encontrados, avaliamos os três melhores classificadores (ET associadas à estilometria, MLP na representação textual e novamente ET na representação textual) no critério geral por meio do teste estatístico de análise da variância simples (*One-way* ANOVA), escolhido por causa da distribuição normal dos exemplos e presença de mais de dois grupos não pareados (Demšar, 2006).

Foram coletadas as métricas dos três classificadores por 30 iterações, usando partições aleatórias. Dado que m é o número de grupos sob análise e n a quantidade de observações, realizamos o cálculo de F e p para todas as bases, usando α=0.05, grau de liberdade no numerador $df1 = m - 1$ e grau de liberdade no denominador $df2 = m - n$. A hipótese nula que defende não haver diferença significativa entre os grupos observados foi rejeitada (Tabela 3).

Tabela 3: Teste estatístico de análise de variância simples (ANOVA *one-way*) a partir da acurácia dos classificadores. O teste também foi aplicado diante dos resultados do classificador geral, composto pela média das três soluções.

Base	$df1$	$df2$	F	p	H_0
Estudantes	2	87	40.5	$3.67 * 10^{-12}$	Rejeitada
Notícias	2	87	858	$5.21 * 10^{-58}$	Rejeitada
Varela	2	87	24855	$1.09 * 10^{-120}$	Rejeitada
Geral	8	269	5539.20	$1.76 * 10^{-286}$	Rejeitada

Com o propósito de explorar modelos híbridos, combinando a estilometria com características textuais, foram criados comitês para classificação combinando os melhores modelos em cada

uma das representações para cada uma das bases de dados. A implementação do comitê se baseou na pilha de classificadores do *Sklearn* e na seleção de características pertinentes para cada classificador. A predição final foi feita combinando a avaliação das predições individuais e uma função de decisão associada a outro classificador - neste caso, a regressão logística.

Contrariando o comportamento esperado, os comitês híbridos não demonstraram benefícios na comparação com os modelos isolados, conforme resultados da Tabela 4. Por esse cenário, pode-se supor que aconteceu uma concorrência entre os classificadores – em lugar da esperada colaboração.

Tabela 4: Resultados finais dos comitês de classificadores híbridos.

Base	Modelo vencedor (acurácia)	Acurácia comitê híbrido
Estudantes	Árvores Extra (0.39)	0.22
Notícias	Árvores Extra (0.94)	0.84
Varela	MLP (0.88)	0.65

4.3 Interpretação dos resultados

O fato de uma solução pautada na representação textual ter alcançado melhores resultados apenas na base de Varela gerou curiosidade devido às observações durante as etapas de análise exploratória e agrupamento com relação a sua composição, que abrange mais tópicos que as demais. Além disso, a base Varela apresenta o maior volume de exemplos dentre as bases estudadas e a maior diversidade de assuntos, uma vez que apresenta textos sobre 10 tópicos diferentes. Consequentemente, seu vocabulário também é superior aos demais. Dadas essas características, é possível que esta base esteja susceptível a sofrer influência do tópico durante a análise de autoria. Ou seja, a distinção dos autores acaba sendo feita pelo conteúdo dos textos e não seu estilo de escrita (Gamon, 2004).

Por causa disso, realizou-se um novo experimento para a base Varela considerando um subconjunto composto por documentos de um único assunto escolhido de maneira aleatória. A metodologia para validação foi a mesma da etapa anterior (validação cruzada estratificada com 3 *folds*). Neste experimento, o ET alcançou 92% de acurácia e 0.96 de AUC, com apenas 1% de diferença para a MLP textual. O experimento provê indícios de que as características de estilo são mais eficazes em domínios específicos do que abrangentes, conforme visto nesse caso e nos anteriores para as bases de estudantes e jornalísticas.

Para aumentar o entendimento sobre os fatores críticos no desenvolvimento da solução, melhorar a compreensão acerca das predições e suportar o entendimento da relevância das características de estilo para verificação de autoria (Goebel et al., 2018), foi realizado um estudo de interpretabilidade dos modelos por meio dos valores Shapley (Shapley, 1953). A técnica foi escolhida por ser capaz de satisfazer diversos axiomas, como eficiência, linearidade, simetria, monotonicidade e proporcionalidade (Sundararajan & Najmi, 2020).

Foram interpretados os melhores classificadores de estilo e gerados gráficos que demonstram as características de maior influência sobre os modelos por meio da biblioteca SHAP[5] capaz de computar os valores Shapley a partir de modelos treinados e apoiar na construção de visualiza-

[5]Biblioteca Python para cálculo de SHAP(*SHapley Additive exPlanations*) - https://github.com/slundberg/shap

ções para interpretação destes valores. Neste experimento, foram avaliados os melhores modelos estilométricos em cada base de dados para identificar as características mais relevantes. As vinte características com melhor desempenho (primeiro quartil), de acordo com os valores Shapley, foram selecionadas para as análises seguintes.

A Figura 11 demonstra as características melhor ranqueadas de cima para baixo, valores superiores indicam que a característica foi mais relevante para o modelo. Cada características é sucedida por pelas iniciais do grupo estilométrico a qual esta pertence entre parêntesis. Na esquerda, tem-se a base de estudantes e na direita a base de notícias. Cada uma das cores representa uma classe (autoria), que é quantificada por meio das barras horizontais, o comprimento total de cada uma das barras indica a importância total da caraterística para o modelo, e as porções das barras coloridas representam o impacto da característica na predição de exemplos daquela classe específica (autoria). Quanto maior a porção colorida da barra horizontal, maior a significância daquela característica na identificação daquele autor e no seu estilo de escrita.

No lado esquerdo da imagem, observamos que para a base de estudantes, as características baseadas em caracteres, lexicais e de riqueza de vocabulário predominam igualmente entre as de maior importância. Já no lado direito, vemos que para a base de notícias também há dominância das características baseadas em caracteres, seguidas pelas de riqueza de vocabulário, lexicais e sintáticas. Para a base Varela, destacam-se as características lexicais e sintáticas, seguidas pelas de domínio da aplicação e baseadas em caracteres ou palavras-chave.

O ordenamento dos grupos de características mais importantes entre as duas primeiras bases e a última são um contraponto interessante, visto que as primeiras possuem domínio restrito, e a outra possui maior diversidade de autores, temas e vocabulário; comprovando a influência do domínio sobre a relevância dos grupos de características.

No lado esquerdo da Figura 11 também observamos grande homogeneidade na importância das características entre os autores, influenciando negativamente para as baixas taxas de acerto observadas para esta base, uma vez que parece mais difícil de distinguir o estilo de escrita destes autores. Também se destacam peculiaridades a partir do cruzamento de informações entre o gráficos e as bases de dados:

- a importância das *stopwords* na maioria dos autores é um indicativo de artigos e preposições no texto, que pode estar associado a esse grupo de estudantes;
- os autores 0 e 10 aparentam possuir um estilo de estruturação de texto através de parágrafos que destoam dos demais;
- a incidência de termos na primeira pessoa e termos no infinitivo (futuro) para os autores 15 e 11 respectivamente, são indicativos de estilos de escrita que permite diferenciá-los.

De maneira geral, observa-se uma maior heterogeneidade na importância das características do lado direito da figura. Isso explica a maior facilidade de diferenciação dos autores durante os experimentos. Pontuamos também:

- o impacto causado pela incidência de vírgulas indica que o autor 5 pode ter um estilo de escrita mais pausado ou sem pausas;

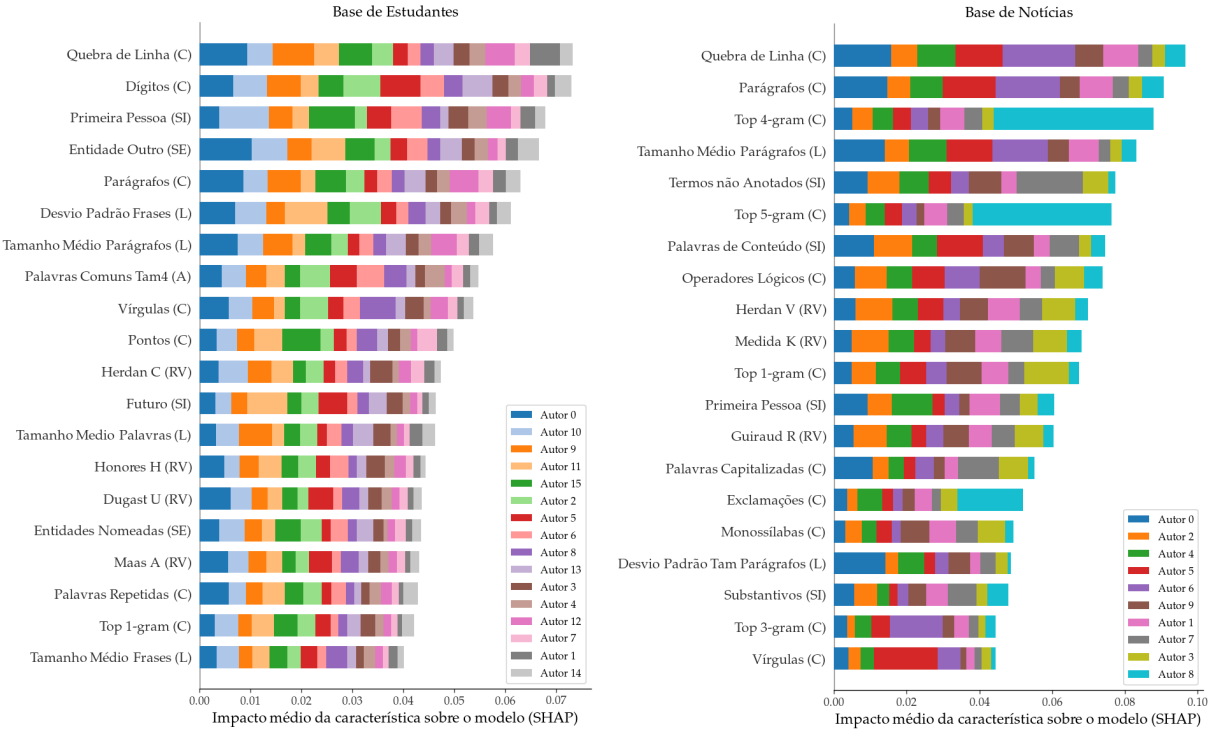

Figura 11: Valores Shapley para base de estudantes e notícias. As características mais importantes estão localizadas acima da imagem. A influência da característica por autor é representada pelas barras coloridas horizontais.

- alta frequência de *top 4-grams* e *5-grams* para o autor 8 são indicativos da presença de sequências de palavras incomuns;
- o autor 7 se destaca pelos termos não etiquetados, que têm relação direta com o *global hapax* e podem ser indicativo de uso de palavras únicas ou estrangeirismos;
- importância das exclamações em textos jornalísticos, que a priori consideramos não ser uma prática comum, mas que pode ter sido introduzida por influência do autor 8.

Com relação à base Varela, devido ao amplo número de autores, torna-se difícil de avaliar visualmente as características mais importantes do ponto de vista individual (Figura 12). Por isso, realizamos um contraste entre os resultados de toda a base (esquerda) e os resultados obtidos a partir de um subconjunto aleatório com todos os trabalhos de 10 autores (300 exemplos), assemelhando-se aos experimentos com bases menores. Na visão macro, é possível confirmar quais grupos de características se destacam globalmente. Já diante do subconjunto descrito acima, foi realizada uma nova análise do ponto de vista individual de cada autor.

Os resultados do subconjunto ratificam observações globais. Individualmente, pode-se elencar:

- o autor 58 se distingue dos demais pelo seu índice de legibilidade (*BR-Flesch*), número de vírgulas e tamanho médio das frases. Uma análise do conteúdo textual dos documentos escritos por esse autor, indica que todos estão contidos no assunto saúde e há uma série

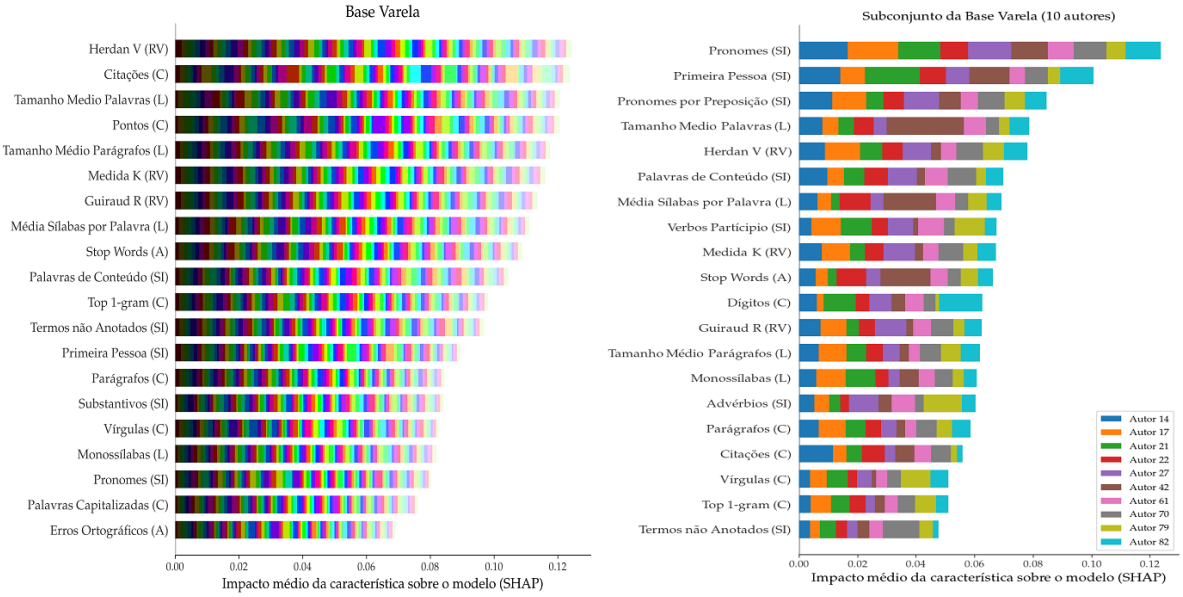

Figura 12: Valores Shapley para base de Varela considerando toda a base do lado esquerdo e um subconjunto com apenas dez autores na base do lado direito, visando facilitar a interpretação dado o número elevado de autores.

de termos técnicos relacionados à medicina, como "arritmia", "hipertireoidismo"e "amiloidose". Concluímos que este autor se distinguiu pelo cunho técnico-científico de seus textos;

- os autores 51 e 82 se sobressaem pelo número de erros ortográficos mas, ao se verificar que o tema principal destes documentos é o turismo, constata-se uma quantidade de palavras estrangeiras superior à dos demais documentos;

- um valor destoante no número de citações do autor 95 nos levou a pensar que seus textos poderiam conter citações literárias ou depoimentos. Entretanto, verificamos que os textos continham entrevistas com autoridades da área da economia.

Como resultado desta análise, inferimos que, independentemente da base, para atividades de atribuição de autoria, o comitê de Árvores Extremamente Aleatórias se beneficia principalmente de características lexicais, baseadas em caracteres, sintáticas e de riqueza de vocabulário. Do ponto de vista individual, foi possível detectar padrões de estilo de escrita relacionados à autoria. Para as bases compostas por textos jornalísticos, nos quais há maior segurança sobre a propriedade intelectual, existe uma maior heterogeneidade na relação entre autoria e características de estilo mais influentes. Já na base de estudantes, onde há uma probabilidade maior de utilização de fontes de informação disponíveis na Web, além da colaboração entre os autores/alunos (Curtis & Tremayne, 2019), as características são bem mais homogêneas.

Na base de estudantes, o *corpora* é composto por vários documentos tratando do mesmo tema e questionamentos, o que acarreta numa restrição de vocabulário, ideologia e formato vinculado às normas institucionais. Por outro lado, as bases jornalísticas, apesar de tratarem de temas relacionados (direito, economia, política), são compostas por documentos escritos por jornalistas e tratam de temas mais abrangentes.

As características de estilo mais relevantes reforçam nossas observações:

- há menor influência de *top-grams* e *hapax* na base de estudantes. Ou seja, não há uma incidência de palavras únicas nos *corpora* suficiente para distinguir os autores, o que pode estar relacionado às limitações já mencionadas;
- forte influência da característica que mensura palavras repetidas, diretamente relacionada ao contexto dos documentos (restrito) e perfil dos autores, que ainda são estudantes e provavelmente possuem habilidades linguísticas menos desenvolvidas que os jornalistas.

5 Conclusões e Trabalhos Futuros

Neste artigo apresentamos uma investigação inovadora ao explorar o uso de características de estilo para identificação de autoria em trabalhos escolares na língua portuguesa. A solução proposta provê um conjunto de características de estilo com efetividade comprovada em português e pode servir de fundamento para trabalhos futuros. As análises, experimentos e observações trazem avanços sobre o uso da estilometria para análise de autoria em trabalhos escolares e atividades pedagógicas.

Neste estudo, constatou-se que o tópico dos documentos analisados exerce grande influência durante as tarefas de atribuição de autoria, corroborando com a utilização de técnicas de classificação que sejam menos afetadas. Dentro da estilometria, ressalta-se a importância da utilização de características não suscetíveis a conteúdo (Halvani, Graner & Regev, 2020). Verificou-se, também, que em bases com alta sobreposição de palavras e limitação de assuntos, as abordagens estilométricas foram superiores às textuais.

Mesmo que a solução proposta não tenha obtido elevadas taxas de acerto na distinção dos estudantes, nenhum outro classificador utilizado durante o estudo foi capaz de superá-la. Para as bases de referência, os resultados alcançados podem não superar o estado da arte, mas solucionam a atividade proposta como uma alternativa eficiente e interpretável. O trabalho apresenta restrições pelo tamanho das bases, que limitaram os experimentos. Em trabalhos futuros, pretende-se incrementar a base de estudantes e avaliar o desempenho da solução diante de conjuntos mais expressivos, aplicar a solução em atividades compartilhadas de análise de autoria, como o PAN[6], e realizar um estudo de campo para avaliar estratégias de inserção da solução desenvolvida em ambientes virtuais de aprendizagem (AVAs).

6 Artigo Premiado Estendido

Esta publicação é uma versão estendida do 1º melhor artigo do Simpósio Brasileiro de Informática na Educação (SBIE) 2021, intitulado “Estudo comparativo entre abordagens estilométricas e textuais para atribuição de autoria em trabalhos escolares”, DOI: 10.5753/sbie.2021.217413

[6]O PAN é um evento anual voltado para resolução de problemas relacionados a análise de autoria. Acesso em: https://pan.webis.de/clef22/pan22-web/index.html

Referências

Agarap, A. F. M. (2018). A neural network architecture combining gated recurrent unit (gru) and support vector machine (svm) for intrusion detection in network traffic data. In *Proceedings of the 2018 10th international conference on machine learning and computing* (pp. 26–30). doi: 10.48550/arXiv:1709.03082 [GS Search]

Aluísio, S., Pelizzoni, J., Marchi, A. R., de Oliveira, L., Manenti, R. & Marquiafável, V. (2003). An account of the challenge of tagging a reference corpus for brazilian portuguese. In *International workshop on computational processing of the portuguese language* (pp. 110–117). doi: 10.1007/3-540-45011-4$_1$7 [GS Search]

Baker, R., Isotani, S. & Carvalho, A. (2011, xx xx). Mineração de Dados Educacionais: Oportunidades para o Brasil. In (Vol. 19, p. xx). SBC. doi: 10.5753/RBIE.2011.19.02.03 [GS Search]

Bevendorff, J., Ghanem, B., Giachanou, A., Kestemont, M., Manjavacas, E., Potthast, M., ... others (2020). Shared tasks on authorship analysis at pan 2020. In *European conference on information retrieval* (pp. 508–516). doi: 10.1007/978-3-030-45442-5$_6$6 [GS Search]

Bezdek, J. C. (2013). *Pattern recognition with fuzzy objective function algorithms*. Springer Science & Business Media. doi: 10.1007/BF00058655 [GS Search]

Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2017). Enriching word vectors with subword information. In (Vol. 5, pp. 135–146). MIT Press. doi: 10.1162/tacl$_{a0}$0051 [GS Search]

Botelho, J. C. & da Silva Martins, M. R. A. (2020). Avaliação da aprendizagem: novas perspectivas para velhos problemas. In (Vol. 2). [GS Search]

Breiman, L. (1996). Bagging predictors. In (Vol. 24, pp. 123–140). Springer. doi: 10.1007/BF00058655 [GS Search]

Breiman, L. (2001). Random forests. In (Vol. 45, pp. 5–32). Springer. doi: 10.1023/A:1010933404324 [GS Search]

Chen, T. & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Chowdhury, G. (2003). Natural language processing. In (Vol. 37, pp. 51–89). doi: 10.1002/aris.1440370103 [GS Search]

Cortez, P. & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. EUROSIS-ETI. [GS Search]

Curtis, G. J. & Tremayne, K. (2019). Is plagiarism really on the rise? results from four 5-yearly surveys. In (pp. 1–11). Taylor Francis. doi: 10.1080/03075079.2019.1707792 [GS Search]

Custódio, J. E. & Paraboni, I. (2021). Stacked authorship attribution of digital texts. In (Vol. 176, p. 114866). Elsevier. doi: 10.1016/j.eswa.2021.114866 [GS Search]

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. In (Vol. 7, pp. 1–30). JMLR. org. [GS Search]

Duarte, V. M. d. N. (2021, 06 10). Sintagma nominal e sintagma verbal.. [Google Search]

Dugast, D. (1979). Vocabulaire et stylistique. In (Vol. 8). Slatkine. [GS Search]

Freitas, C., Carvalho, P., Gonçalo Oliveira, H., Mota, C., Santos, D. et al. (2010). Second harem: advancing the state of the art of named entity recognition in portuguese. In *Proceedings of the international conference on language resources and evaluation (lrec 2010)(valletta 17-23 may de 2010) european language resources association.* [GS Search]

Gamon, M. (2004). Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Coling 2004: Proceedings of the 20th international conference on computational linguistics* (pp. 611–617). doi: 10.3115/1220355.1220443 [GS Search]

Geurts, P., Ernst, D. & Wehenkel, L. (2006). Extremely randomized trees. In (Vol. 63, pp. 3–42). Springer. doi: 10.1007/s10994-006-6226-1 [GS Search]

Gillam, L. & Vartapetiance, A. (2012). Quite simple approaches for authorship attribution, intrinsic plagiarism detection and sexual predator identification.. [GS Search]

Goebel, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., Stumpf, S., ... Holzinger, A. (2018). Explainable ai: the new 42? In *International cross-domain conference for machine learning and knowledge extraction* (pp. 295–303). doi: 10.1007/978-3-319-99740-7$_2$1 [GS Search]

Goldberg, Y. (2017). Neural network methods for natural language processing. In (Vol. 10, pp. 1–309). Morgan & Claypool Publishers. doi: 10.2200/S00762ED1V01Y201703HLT037 [GS Search]

Halvani, O., Graner, L. & Regev, R. (2020). A step towards interpretable authorship verification.. doi: 10.48550/arXiv.2006.12418 [GS Search]

Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Silva, J., Aluísio, S. et al. (2017, oct). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Proceedings of the 11th Brazilian symposium in information and human language technology* (pp. 122–131). Uberlândia, Brazil: Sociedade Brasileira de Computação. doi: 10.48550/arXiv.1708.06025 [GS Search]

Honnibal, M. & Montani, I. (2017). spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing.

Honoré, A. (1979). Some simple measures of richness of vocabulary. In (Vol. 7, pp. 172–177). [GS Search]

Jang, B., Kim, I. & Kim, J. W. (2019). Word2vec convolutional neural networks for classification of news articles and tweets. In (Vol. 14, p. e0220976). Public Library of Science San Francisco, CA USA. doi: 10.1371/journal.pone.0220976 [GS Search]

Juola, P. (2008). *Authorship attribution* (Vol. 3). Now Publishers Inc. doi: 10.1561/1500000005 [GS Search]

Khonji, M., Iraqi, Y. & Jones, A. (2015). An evaluation of authorship attribution using random forests. In *2015 international conference on information and communication technology research (ictrc)* (pp. 68–71). doi: 10.1109/ICTRC.2015.7156423 [GS Search]

Kocev, D., Vens, C., Struyf, J. & Džeroski, S. (2013). Tree ensembles for predicting structured outputs. In (Vol. 46, pp. 817–833). Elsevier. doi: 10.1016/j.patcog.2012.09.023 [GS Search]

Kumawat, D. & Jain, V. (2015). Pos tagging approaches: A comparison. In (Vol. 118). Citeseer. doi: 10.5120/20752-3148 [GS Search]

Maitra, P., Ghosh, S. & Das, D. (2016). Authorship verification-an approach based on random forest.. doi: 10.48550/arXiv.1607.08885 [GS Search]

Martins, T. B., Ghiraldelo, C. M., Nunes, M. d. G. V., de Oliveira Junior, O. N. et al. (1996). Readability formulas applied to textbooks in brazilian portuguese. Icmsc-Usp. [GS Search]

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space.. doi: 10.48550/arXiv.1301.3781 [GS Search]

Neal, T., Sundararajan, K., Fatima, A., Yan, Y., Xiang, Y. & Woodard, D. (2017). Surveying stylometry techniques and applications. In (Vol. 50, pp. 1–36). ACM New York, NY, USA.

doi: 10.1145/3132039 [GS Search]

Pacheco, M. L., Fernandes, K. & Porco, A. (2015). Random forest with increased generalization: A universal background approach for authorship verification. In *Clef (working notes)*. [GS Search]

Pardo, T. A. S. & Nunes, M. d. G. V. (2003). A construção de um corpus de textos científicos em português do brasil e sua marcação retórica.. [GS Search]

Peng, J., Choo, R. K.-K. & Ashman, H. (2016). Astroturfing detection in social media: Using binary n-gram analysis for authorship attribution. In *2016 ieee trustcom/bigdatase/ispa* (pp. 121–128). doi: 10.1109/TrustCom.2016.0054 [GS Search]

Pires, A. R. O. (2017). *Named entity extraction from portuguese web text*. Unpublished master's thesis, Faculdade de Engenharia da Universidade Do Porto. [GS Search]

Ramshaw, L. A. & Marcus, M. P. (1999). Text chunking using transformation-based learning. In *Natural language processing using very large corpora* (pp. 157–176). Springer. doi: 10.1007/978-94-017-2390-9$_1$0 [GS Search]

Santos, D. & Zanchettin, C. (2021). Estudo comparativo entre abordagens estilométricas e textuais para atribuição de autoria em trabalhos escolares. In *Anais do xxxii simpósio brasileiro de informática na educação* (pp. 760–772). Porto Alegre, RS, Brasil: SBC. doi: 10.5753/sbie.2021.217413 [GS Search]

Scarton, C. E. & Aluísio, S. M. (2010). Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh-metrix para o português. In (Vol. 2, pp. 45–61). [GS Search]

Schapire, R. E., Singer, Y. & Singhal, A. (1998). Boosting and rocchio applied to text filtering. In *Proceedings of the 21st annual international acm sigir conference on research and development in information retrieval* (pp. 215–223). doi: 10.1145/290941.290996 [GS Search]

Shapley, L. S. (1953). A value for n-person games. In (Vol. 2, pp. 307–317). doi: 10.1515/9781400881970-018 [GS Search]

Shrestha, P., Sierra, S., González, F. A., Montes, M., Rosso, P. & Solorio, T. (2017). Convolutional neural networks for authorship attribution of short texts. In *Proceedings of the 15th conference of the european chapter of the association for computational linguistics: Volume 2, short papers* (pp. 669–674). [GS Search]

Silva, D. d. C. (2011). Algoritmos de processamento da linguagem e síntese de voz com emoções aplicados a um conversor texto-fala baseado em hmm. *Doutorado, Programa de Engenharia Elétrica, Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia (COPPE/UFRJ), Rio de Janeiro*. [GS Search]

Singh, S. & Remenyi, D. (2016). Plagiarism and ghostwriting: The rise in academic misconduct. In (Vol. 112, pp. 1–7). Academy of Science of South Africa. doi: 10.17159/sajs.2016/20150300 [GS Search]

Soares, F., Yamashita, G. H. & Anzanello, M. J. (2018). A parallel corpus of theses and dissertations abstracts. In *International conference on computational processing of the portuguese language* (pp. 345–352). doi: 10.1007/978-3-319-99722-3$_3$5 [GS Search]

Stamatatos, E. (2009). A survey of modern authorship attribution methods. In (Vol. 60, pp. 538–556). Wiley Online Library. doi: 10.1002/asi.21001 [GS Search]

Sundararajan, M. & Najmi, A. (2020, 13–18 Jul). The many shapley values for model explanation. In H. D. III & A. Singh (Eds.), *Proceedings of the 37th international conference on machine*

learning (Vol. 119, pp. 9269–9278). PMLR. [GS Search]

Tarekegn, A. N., Giacobini, M. & Michalak, K. (2021). A review of methods for imbalanced multi-label classification. In (Vol. 118, p. 107965). Elsevier. doi: 10.1016/j.patcog.2021.107965 [GS Search]

Tempestt, N., Kalaivani Sundararajan, A. F., Yan, Y., Xiang, Y., Woodard, D. et al. (2017). Surveying stylometry techniques and applications. In (Vol. 50). doi: 10.1145/3132039 [GS Search]

Thinsungnoena, T., Kaoungkub, N., Durongdumronchaib, P., Kerdprasopb, K., Kerdprasopb, N. et al. (2015). The clustering validity with silhouette and sum of squared errors. In (Vol. 3). doi: 10.12792/iciae2015.012 [GS Search]

Tweedie, F. J. & Baayen, R. H. (1998). How variable may a constant be? measures of lexical richness in perspective. In (Vol. 32, pp. 323–352). Springer. doi: 10.1023/A:1001749303137 [GS Search]

Van der Maaten, L. & Hinton, G. (2008). Visualizing data using t-sne. In (Vol. 9). [GS Search]

Varela, P. J., Albonico, M., Justino, E. J. R., Bortolozzi, F. et al. (2018). A computational approach for authorship attribution on multiple languages. In *2018 international joint conference on neural networks (ijcnn)* (p. 1-8). doi: 10.1109/IJCNN.2018.8489704 [GS Search]

Weisberg, S. (2001). Yeo-johnson power transformations. In (Vol. 1, p. 2003). [GS Search]

Werneck, V. R. (2006). Sobre o processo de construção do conhecimento: o papel do ensino e da pesquisa. In (Vol. 14, pp. 173–196). SciELO Brasil. doi: 10.1590/S0104-40362006000200003 [GS Search]

Yang, M., Chen, X., Tu, W., Lu, Z., Zhu, J. & Qu, Q. (2018). A topic drift model for authorship attribution. In (Vol. 273, pp. 133–140). Elsevier. doi: 10.1016/j.neucom.2017.08.022 [GS Search]

Yule, G. (1944). The statistical study of literary vocabulary. cambridge, cambridge [eng.]. University Press. Journal of the Royal Statistical Society. doi: 10.2307/2981280 [GS Search]